Advances in Applied Mathematics 应用数学进展, 2017, 6(5), 692-697
Published Online August 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.65082

Hans汉斯

Exploration to Achieve by MATLAB Modeling Linear Regression Processing, Processing Environmental Monitoring Data and Comparison with Excel Modeling

Jialiang Sun

School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou Henan
Email: 13027522521@163.com

Received: Jul. 22nd, 2017; accepted: Aug. 11th, 2017; published: Aug. 17th, 2017

Abstract

This thesis, using linear regression theory and MATLAB mathematical software, explores the solution of the standard curve in the processing of environmental monitoring data. Compared with the Excel modeling method, this modeling method for solving the linear regression equation of standard curve, can make it's error smaller and make it more accurate and simple to operate, and can improve the work efficiency. It can be fully used in the part of the chemical work.

Keywords

Environmental Monitoring, Standard Curve, Linear Regression, MATLAB

MATLAB建模实现线性回归处理环境监测数据的探索及与EXCEL建模的对比

孙嘉良

华北水利水电大学数学与统计学院，河南 郑州
Email: 13027522521@163.com

收稿日期：2017年7月22日；录用日期：2017年8月11日；发布日期：2017年8月17日

文章引用：孙嘉良. MATLAB 建模实现线性回归处理环境监测数据的探索及与 EXCEL 建模的对比[J]. 应用数学进展, 2017, 6(5): 692-697. DOI: 10.12677/aam.2017.65082

摘　要

本文以线性回归理论为基础，运用MATLAB数学软件进行建模，针对环境监测数据处理中标准曲线法的求解问题进行探究。该建模方法所求得标准曲线的线性回归方程较Excel建模方法误差小，更为准确，可提高工作效率，在科研工作中可以得到一定应用。

关键词

环境监测，标准曲线，线性回归，MATLAB

1. 引言

对于环境监测数据处理时求解标准曲线问题，其目的是利用得到的标准曲线计算待测物质的含量，以求精确评价所测水体环境或者采取合适的水处理工艺。

对此问题目前化学工作者运用基于 Excel 程序进行求解，如白云做过 c++程序求解的探究[1]，其求解误差小，但是编程过程复杂。刘作云等运用最小二乘法基于 Excel 进行求解[2] [3] [4]，但得到的结果误差较大。求解标准曲线，对于化学工作者来说需要具有一定的数据处理能力以及软件操作能力，除此之外求得的标准曲线误差大小也是一直被关注与关心的。本文采用基于 MATLAB 的操作，将线性回归方法应用于光谱分析中，可以得到误差更小的标准曲线方程。并且本文对求解结果进行了 p 检验和 F 检验以及相关系数 R 的计算，结果都很好的符合要求。

2. 理论基础

线性回归利用线性回归方程的最小平方函数，对自变量和因变量之间的关系进行建模。国家规定的标准曲线做法是测定已知待测物质含量的标准试剂的吸光度值，得到一些离散点，求解通过这些点的一条直线。

由光吸收基本定律可知当实验条件满足要求时，待测物质的含量与吸光度呈线性关系，因而线性回归可运用于标准曲线求解。二者满足线性关系的前提下，影响吸光度值的条件中仅物质含量一个变量，因此采用一元线性回归分析来进行二者关系的建模。本文以吸光度值x为自变量，待测物含量y为因变量，建立一元线性回归模型：

$$y=\beta x+\varepsilon \quad [5]$$

步骤一：以国家规定的标准溶液浓度系列作为矩阵，实验得到对应的吸光度值矩阵，即

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}=\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}+\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (*)$$

假定矩阵 $\varepsilon \sim \mathrm{N}\left(0,\sigma^2\right)$，$\beta$ 为需要顾及的系数向量，ε 为随机误差向量[6]。

步骤二：对于准备好的数学模型，将其运用到 MATLAB 软件的操作当中，编程调用 MATLAB 中的 regress 函数作一元线性回归：

$$b = \text{regress}(Y, X)$$

返回线性回归方程中系数向量 β 的估计值 b，regress 函数将 Y 或 X 中不确定的数据作为缺失数据而忽略它们，使得误差会变小。最终可以得到回归方程系数的估计值。

步骤三：进行显著性检验和误差分析。通过得到的显著性检验相关参数，与临界值比较，采用残差进行误差分析，残差是衡量不确定性的指标，残差大小可以衡量预测的准确性，残差越大表示预测越不准确，残差与数据本身的分布特性和回归方程的选择有关。

$$\text{残差的均值} = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)}{n}$$

其中 y_i 为因变量的真实值，$\widehat{y_i}$ 为估计值。

3. 案例分析

3.1. 案例一

以“紫外分光光度法”测总氮含量中标准曲线求解为例，在满足定量分析要点的基础上，计算出标准试剂的浓度和扣除空白后对应的校正吸光度值(以下直接称吸光度值)，得到标准溶液的总氮含量(Y)和吸光度值(X)，如下表 1：

构造出表中对应的线性回归矩阵算法如下：

$$\begin{pmatrix} 0.00 \\ 0.005 \\ 0.02 \\ 0.03 \\ 0.05 \\ 0.07 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} \begin{pmatrix} 0.000 \\ 0.048 \\ 0.208 \\ 0.304 \\ 0.506 \\ 0.716 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

MATLAB 运行结果(表 2~4，图 1)

以上即利用 MATLAB 数学软件求得的结果及各统计参数，由表 2 可知求得回归方程(标准曲线方程)为 $Y = 0.0989X - 8.1764e-05$。

显著性检验

我们将原假设与对立假设分别为

$$H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0$$

1) 由表 4 检验 p 值为 2.6847e−10，远小于 0.01，可知显著性水平 $\alpha = 0.01$ 下应拒绝原假设 H_0，可认为吸光度值与浓度的线性关系是显著的；

2) 对于 F 统计量的观测值与临界值 $F_{\alpha(1,n-2)}$，当 $F \geq F_{\alpha(1,n-2)}$ 时，拒绝原假设，认为 Y 与 X 的线性关系是显著的[7]；反之，则接受原假设，认为 Y 与 X 的线性关系是不显著的。在表 4 中 F 统计量的观测值 $2.1862e+04 > F_{0.01(1,4)} = 21.20$，可知吸光度值与浓度有显著的线性关系。

与 Excel 建模方法[2]进行对比

将案例一的数据运用 Excel 建模方法最终得到线性回归方程为：$Y = 0.09886745X + 0.00008176$，求得其残差的均值为−5.46429e−09；在表 3 中得到 Y 的残差分析，可以看出通过 MATLAB 求得的线性回

Table 1. Total nitrogen standard reagent concentration and calibration absorbance value
表 1. 总氮标准试剂浓度与校正吸光度值

X	0.000	0.048	0.208	0.304	0.506	0.716
Y (mg)	0.00	0.005	0.02	0.03	0.05	0.07

Table 2. The estimate of the coefficients, 95% confidence upper limit and 95% lower confidence limit
表 2. 系数的估计值与其置信上下限

系数的估计值	估计值的 95%置信下限	估计值的 95%置信上限
−8.1764e−05	−8.9099e−04	7.2746e−04
0.0989	0.0971	0.1006

Table 3. The real value of Y corresponds to the estimate, residual and 95% confidence interval
表 3. Y 的真实值对应的估计值、残差及 95%置信区间

Y 的真实值	Y 的估计值	残差	残差的 95%置信下限	残差的 95%置信上限
0	−8.1764e−05	8.1764e−05	−0.0011	0.0013
0.0050	0.0047	3.3613e−04	−8.1348e−04	0.0015
0.0200	0.0205	−4.8266e−04	−0.0017	7.0416e−04
0.0300	0.0300	2.6060e−05	−0.0013	0.0014
0.0500	0.0499	5.4836e−05	−0.0013	0.0014
0.0500	−8.1764e−05	8.1764e−05	−0.0011	0.0013
0.0700	0.0047	3.3613e−04	−8.1348e−04	0.0015
0.0800	0.0205	−4.8266e−04	−0.0017	7.0416e−04

Table 4. F statistic observation value, p value, residual mean and R^2
表 4. F 统计量观测值、p 值、残差均值及 R^2

F 统计量观测值	检验的 p 值	残差的均值	R^2
2.1862e+04	2.6847e−10	0	0.9998

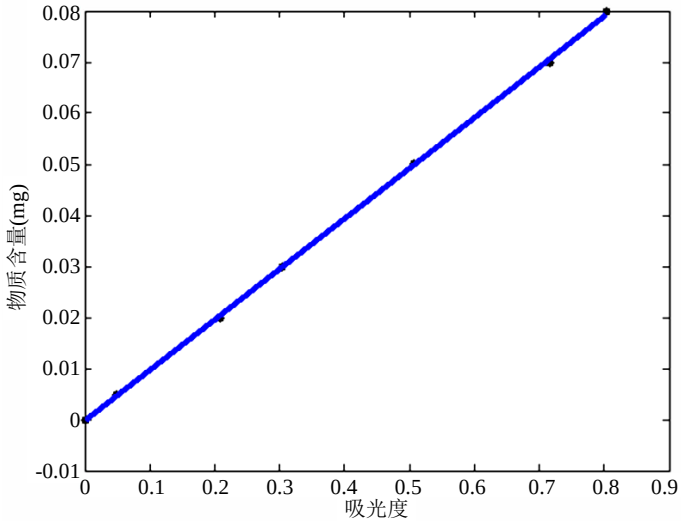

Figure 1. The standard curve fitting of total nitrogen
图 1. 总氮的标准曲线拟合图

归方程的残差的均值(表 4)在同样的精确度下为 0，说明 MATLAB 建模较 Excel 建模的误差小，准确度高。

3.2. 案例二

下面以硝氮含量测定实验数据对所建立的模型进行检验，吸光度值 X 的数据如标准试剂中硝氮的含量 Y 下表 5：

MATLAB 程序运行结果：(表 6~8，图 2)

由表 6，回归方程(标准曲线方程)为 $Y = 205.0419X - 0.0419$。

显著性检验

由表 8 得出 p 值为 2.7987e−09 远远小于 0.01，F 统计量的观测值为 $4.6299\text{e}+04 > F_{0.01(1,4)} = 21.20$，可知该组数据的吸光度值与浓度有显著的线性关系。

与 Excel 建模方法[2]进行对比

将案例二的数据运用 Excel 建模方法最终得到线性回归方程为：$Y = 205.04186X + 0.04191$，求得其残差的均值为−1.76667e−07；在表 7 中得到 Y 的残差分析，可以看出通过 MATLAB 求得的线性回归方程的残差的均值(表 8)为−1.38778e−17，可以看 $\left|-1.38778\text{e}-17\right| \ll \left|-1.76667\text{e}-07\right|$。

Table 5. Nitrate nitrogen standard reagent and calibration absorbance value
表 5. 硝氮标准试剂浓度与校正吸光度值

X	0.000	0.058	0.124	0.245	0.365	0.487
Y (μg)	0.000	12.50	25.00	50.00	75.00	100.00

Table 6. The estimate of the coefficients, 95% confidence upper limit and 95% lower confidence limit
表 6. 系数的估计值与其置信上下限

系数的估计值	估计值的 95%置信下限	估计值的 95%置信上限
0.0419	−0.6820	0.7658
205.0419	202.3961	207.6876

Table 7. The real value of Y corresponds to the estimate, residual and 95% confidence interval
表 7. Y 的真实值对应的估计值、残差及 95%置信区间

Y 的真实值	Y 的估计值	残差	残差的 95%置信下限	残差的 95%置信下限
0	0.0419	−0.0419	−1.0137	0.9298
12.5	11.9343	0.5657	−0.0053	1.1366
25	25.4671	−0.4671	−1.3262	0.3920
50	50.2772	−0.2772	−1.3570	0.8027
75	74.8822	0.1178	−0.9415	1.1771
100	99.8973	0.1027	−0.7010	0.9064

Table 8. F statistic observation value, p value, residual mean and R^2
表 8. F 统计量观测值、p 值、残差均值及 R^2

F 统计量的观测值	检验的 p 值	残差的均值	R^2
4.6299e+04	2.7987e−09	−1.38778e−17	0.9999

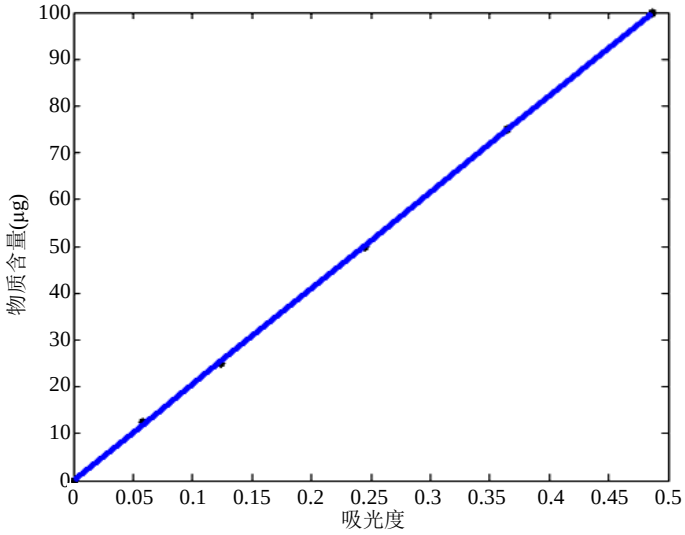

Figure 2. The standard curve fitting of nitrous nitrogen
图 2. 硝氮的标准曲线拟合图

结果分析

两个案例基于不同的化学实验，测定的是不同的物质，标准物质的含量、吸光度值也不同，通过最终的实验结果可以看出，本文建立的模型可以适用于不同要求标准曲线方程的求解，且通过与 Excel 建模方法对比，可知 MATLAB 建模的误差较小，得到的回归方程准确度更高。

4. 结论

对于标准曲线的求解问题，一直未有基于 MATLAB 的线性回归分析方法。本文通过探索 MATLAB 建模实现线性回归，可快速的得到物质含量与吸光度值的标准曲线方程，同时和 Excel 建模对比，减小了误差，可以得到更准确的待测物质含量，从而得对水环境质量做出更为准确的评估。

参考文献 (References)

[1] 白云. 利用 C 语言学习一元线性回归处理[J]. 湖北科技学院学报, 2015, 35(10): 195-197.

[2] 刘作云. 最小二乘法处理环境监测数据及 Excel 建模探索[J]. 湖南生态科学学报, 2015, 2(3): 26-30.

[3] 单文坡, 卢海霞, 郑辉. Excel 和 Origin 在环境监测数据处理中的应用[J]. 石家庄职业技术学院, 2008, 20(2): 58-60, 67.

[4] 吴慧璇, 林宙峰, 郭益军. Excel 软件在环境监测数据处理中的应用[J]. 科教文汇(上旬刊), 2007(7): 202.

[5] 谢中华, 李国栋, 刘焕进, 吴鹏, 郑志勇. MATLAB 从零到进阶[M]. 北京: 北京航空航天大学出版社, 2012: 377-379.

[6] 林彬. 多元线性回归分析及其应用[J]. 中国科技信息, 2010(9): 60-61.

[7] 任建英. 一元线性回归分析及其应用[J]. 才智, 2012(22): 116-117.